View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by OpenCommons at University of Connecticut

University of Connecticut
OpenCommons@UConn

Master's Theses

University of Connecticut Graduate School

5-9-2014

Infants May Be Sensitive to Asynchronous Audiovisual Speech

Kathleen E. Shaw
kathleen.shaw@uconn.edu

Recommended Citation

Shaw, Kathleen E., "Infants May Be Sensitive to Asynchronous Audiovisual Speech" (2014). *Master's Theses*. 609.
https://opencommons.uconn.edu/gs_theses/609

This work is brought to you for free and open access by the University of Connecticut Graduate School at OpenCommons@UConn. It has been accepted for inclusion in Master's Theses by an authorized administrator of OpenCommons@UConn. For more information, please contact opencommons@uconn.edu.

Infants May Be Sensitive to Asynchronous Audiovisual Speech

Kathleen Elizabeth Shaw

M. A. I. S., Oregon State University, 2011

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Arts

At the

University of Connecticut

2014

APPROVAL PAGE

Masters of Arts Thesis

Infants May Be Sensitive to Asynchronous Audiovisual Speech

Presented by

Kathleen Elizabeth Shaw, M. A. I. S.

Major Advisor_____
Heather Bortfeld

Associate Advisor_____
Nicole Landi

Associate Advisor_____
Rachel Theodore

University of Connecticut

2014

Infants May Be Sensitive to Asynchronous Audiovisual Speech

Infants undergo a remarkable transformation in their perception of language across the first year of life. At birth, they are considered universal speech perceivers, able to discriminate phonetic contrasts of all languages despite a lack of exposure (Werker & Tees, 1984). Around 6-months, they can rely on distributional frequencies of speech sounds to segment words from speech streams (Maye, Werker, & Gerken, 2002) and also become expert perceivers to vowels of their native sounds at the expense of non-native vowels (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992). Frequently, developmental investigations of these phenomena have approached speech perception as an auditory-only process, focusing on infant sensitivity to the acoustic signals of spoken language. However, speech perception is inherently a multimodal process, with information provided by both the auditory and visual modalities to the developing language learner.

The audiovisual speech signal, composed of both visual and auditory information, provides relatively stable, tightly correlated multimodal cues that infants can utilize when perceiving spoken language. Both sensory streams share amodal voicing onsets and offsets, intensities, amplitude contours, durations, and rhythmic patterning which reliably co-occur and facilitate audiovisual speech integration (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). However, our understanding of infant sensitivity to these coupled cues and when these discriminative abilities develop is still an area of controversy within the developmental literature. How is it that infants bind the appropriate sights to sounds in their environment and is it a faculty available at birth or a perceptual process constructed as they gain experience with cause-effect relations of the world?

As young infants are unable to provide explicit responses due to their undeveloped language and fine motor skills, developmental researchers have had to find implicit measures of infant sensitivity to audiovisual speech perception, including looking time paradigms (e.g., Desjardins & Werker, 2004), electrophysiological approaches (e.g., Kopp & Dietrich, 2013), and neurophysiological techniques (e.g., Watanabe et al., 2013). The current study examined infant sensitivity to low-level cues across modalities when perceiving speech using a modified intermodal preferential looking paradigm, a technique commonly used to examine infant perceptual discrimination.

The intermodal preferential looking paradigm (IPLP) has become a staple of developmental language studies and has shed new light on a variety of infant comprehension measures, including word learning (Pruden, Hirsh-Pasek, Golinkoff, & Hennon, 2006), verb acquisition (Naigles, 1990), vocabulary knowledge (Bergelson & Swingley, 2012), and word order (Golinkoff, Hirsh-Pasek, Cauley, & Gordon, 1987). Although there are many modifications dependent upon the aims of the particular question of interest, the typical IPLP consists of the infants sitting in front of two computer monitors flanking a centrally-located speaker. At-test, infants are presented with two different visual displays simultaneously while audio is presented. One of the visual displays (the *target* or *match*) will be congruent with the audio signal (e.g., a picture of an apple while the audio presents “apple”) while the other visual display (the *distractor* or *non-match*) will be incongruent with the audio signal (e.g., a picture of a banana while the audio presents “apple”). During presentation, infant visual gaze is recorded to both the match and non-match screens and comprehension is measured by whether they show a preference for the screen that matched the auditory stimulus over the screen that did not. The

IPLP has proven to be an invaluable method of investigating multimodal perception and, importantly, audiovisual speech perception.

For instance, Patterson and Werker (2003) investigated whether infants as young as two-months-old are capable of matching an auditory phoneme to the appropriate articulating face in an IPLP. Visual presentation was either of /i/ or /a/ and was produced by either a male or female actor. Regardless of gender, infants were able to match the auditory stimulus to the appropriate visual presentation. The early aptitude of correctly matching phonetic information across two modalities suggests that there is an attentional mechanism facilitating phonetic discrimination and speech perception and that is in on-line even early in development. However, it is still unclear which cues in the auditory and visual signals are driving this multisensory perceptual experience and available to the very young infant.

To the perceiver, there are three types of cues that are available in the audiovisual speech signal – spectral cues, temporal cues, and phonetic cues (Baart, Vroomen, Shaw, & Bortfeld 2014). Spectral cues refer to the energy and temporal correlations contained within the auditory signal and how it is produced by the vocal tract. For instance, the vowels /a/ and /i/ differ in their height, backness and first formants (F_1) with /a/ being a low, back vowel and /i/ being a high, front vowel. Together, these two cues and the vocal tract, along with other spectral properties, help shape the acoustic signal and the energy contained therein. Temporal cues refer to characteristics across the two modalities that are coupled in time. For instance, the displacement, velocity, and acceleration of the lips coming together (known as a bilabial place of articulation) directly shape the acoustic signal and provide visual cues to the cause-effect relationship of the articulators and audible speech. Finally, phonetic cues, or spectral and temporal correspondences that delineate a /b/ from a /p/ from an /m/, despite all including bilabial closure, are also available

in audiovisual speech perception. Regarding Patterson and Werker's (2003) findings, it is unclear which of these cue types may be facilitating two-month-olds in matching the auditory vowel to the appropriate visual display. In all presentations, the two signals were presented simultaneously and in naturally spoken speech, providing reliable spectral and temporal cues to their congruency. However, accessibility to phonetic cues is debatable, as it is difficult to disentangle whether infants were simply relying on the previously mentioned low-level cues or were applying phonetic knowledge while processing the audiovisual signals.

To better assess which cue types infants might be relying on when perceiving speech, Kuhl and colleagues (1982, 1984, & 1991) investigated whether modulating the spectral information of the acoustic signal impaired infant audiovisual speech processing. In a series of IPLP designs, 4- to 5-month-old infants were presented with audiovisual displays similar to Patterson and Werker (2003) with the vowels being reduced in their spectral properties and replaced by either pure tones, tones that matched the F_1s of the vowels, or three-tone vowel analogues (matching F_1, F_2, and F_3 of the naturally spoken vowels) (Kuhl & Meltzoff, 1984; Kuhl, Williams, & Meltzoff, 1991, respectively). When given the natural acoustic speech signal, infants matched the auditory vowels to the appropriate articulating face. However, across all three spectral manipulations, they failed to preferentially attend to the matching face over the mismatching face. These findings would suggest that temporal correlations between the signals are not enough for infants to audiovisually integrate the two modalities. Spectral information, particularly those that go above and beyond the basic formant frequencies of the vowels is necessary to combine heard and seen speech. As a result of their findings, Kuhl and colleagues suggest that the phonetic aspects of the speech signal drive multimodal integration. They purport that infants need to be given a natural speech stimulus to process cross-modal relationships of the

speech sound and the articulators and that audiovisual speech perception is a holistic process, with infants being relatively insensitive to the finer cues available.

In contrast to Kuhl and colleagues (1982, 1984, & 1991), recent studies have suggested that infants are sensitive to spectral cues in the speech signal and the visual articulators. Pena, Mehler, and Nespor (2011), found that infants match the frequencies of auditory vowels to shape sizes that mimic the natural movements of the mouth during production. In this case, the visual speech cues would be degraded, as they share the general size proportions of the mouth but are devoid of articulatory information. When infants were presented with a high frequency sound similar to /e/, they were more likely to attend to smaller objects in an IPLP. When presented with a lower frequency sound similar to /a/, they performed the opposite pattern, focusing on larger shapes. Although limited in its inferences to natural visual speech, Pena et al. (2011) suggest that their findings demonstrate infant sensitivity to how the mouth moves and the resulting acoustic signal, which might in turn inform an amodal representation of acoustic sound and visual size as a reflection of experience with natural speech.

Another recent study suggesting that infants may be able to integrate audiovisual cues under degraded speech conditions compared audiovisual matching with either natural speech or sine wave speech (Baart et al., 2014). Sine wave speech (SWS) is natural speech that has been synthetically reduced to three sinusoids that replicate the first three formants in frequency and amplitude (Remez, Rubin, Pisoni, & Carrell, 1981). Unlike natural speech signals, it consists of degraded phonetic cues yet keeps the spectrotemporal qualities of natural speech. Studies have found that adults have difficulty in recognizing the underlying phonetic content of SWS unless they have been trained to recognize the SWS as language, a phenomenon known as being put in "speech-mode" (Eskelund, Tuomainen, & Andersen, 2011; Remez et al., 1981; Tuomainen,

Andersen, Tippana, & Sams, 2005; Vroomen & Baart, 2009; Vroomen & Stekelenburg, 2011). When presented with an IPLP display and either natural speech or SWS, infants matched the auditory word to the articulating mouth equally well for both speech types. This suggests that they can rely on the temporal dynamics and limited spectral cues available within the reduced signal, contrary to Kuhl et al.'s (1984, 1991) findings with similar, but shorter, vowels. When adults were presented with the same stimuli and asked to identify which screen match the auditory signal, they performed significantly worse for SWS auditory token than natural speech tokens, which may suggest that their heightened experience with phonetic cues actually hindered performance – by acclimating to relying on natural phonetic properties of speech, they failed to recognize the cues available to them across the two modalities. An area of inquiry is how reliance on different cues develops as a reflection of experience and positing theories on what may or may not be available or salient to the perceiver based on developmental trajectories of audiovisual perception.

There is a severe disconnect between the developmental and adult literature regarding audiovisual perception and sensitivity to the fine-grained cues across modalities. For instance, children have been shown to be less sensitive to incongruencies between visual and auditory information than adults, suggesting that the development of audiovisual integration may be protracted or inconsistent through adolescence (Hillock, Powers, & Wallace, 2011; Innes-Brown et al., 2011; McGurk & McDonald, 1971; Pons, Teixido, Garcia-Morera, & Navarra, 2012). In particular, children have been shown to erroneously provide simultaneity judgments when an auditory signal preceded a visual stimulus, suggesting that they are likely to perceive the asynchronous events as a unified percept (Hillock et al., 2011). Even within the infant literature, there have been suggestions that infants are susceptible to perceiving asynchronous relationships

as a unified whole, even when the incongruencies between the signals are particularly salient to adults. Below we briefly discuss two studies including infants that demonstrate their susceptibility to erroneous audiovisual integration and purport that short-term experience may drive perceptual discrimination of temporal correlations.

Pons, Teixido, Garcia-Morera, and Navarra (2012) investigated whether cues to asynchrony may help young infants in recognizing the incongruencies between audio and visual information. In their study, 6-month-old infants were presented with an IPLP consisting of two bouncing ball presentations. For one of the display sides, the ball was accompanied by an impact noise that was synchronous with the ball hitting either the sides or the floor of the display. For the opposite display, the impact sound preceded the ball reaching the screen frame by 500ms, an auditory-lead that children have been found to be insensitive to in child-adult comparison studies (e.g., Hillock et al., 2011). During the experimental session, infants completed three phases – an initial preference phase to assess discriminatory abilities between the two screens, an exposure-to-asynchrony phase, in which they were only presented with the asynchronous ball display, and finally a test phase, identical to the initial preference phase. Comparing looking times across the initial preference phase and the test phase, Pons et al. found that infants were *more* likely to prefer the asynchronous display at test than during the initial preference phase. They assert that by highlighting the asynchronous nature of the visual ball bounce and the auditory signal of the impact sound, that infants were capable of recognizing the difference which guided their looking preferences at test. These findings suggest that infants *can* be sensitive to the temporal congruency (or incongruency) of an audiovisual presentation, however, the oddball nature of the task must be made salient for discrimination. Importantly, this study did not involve speech or speech stimuli, which would naturally differ based on the complex spectral information of the

multimodal signal. Indeed, studies have shown that infants prefer biological motion over audiovisual synchrony (Falck-Ytter, Bakker, & von Hofsten, 2011) among other social stimuli, which may lead to them being more in-tune to violations of socially-relevant signals. In that case, one question to consider is whether infants also insensitive to temporal asynchrony in audiovisual speech if the differences have not been made salient to them.

Lewkowicz (2010) conducted a similar study investigating whether infant sensitivity to asynchronous audiovisual displays is dependent upon highlighting asynchronous relations first. Infants between 4- and 10-months old were habituated to either a synchronous presentation, in which an auditory /ba/ was simultaneously presented with a face articulating /ba/ or an asynchronous presentation, in which an auditory /ba/ was presented 666ms prior to the visual /ba/. At test, infants were exposed in a between-subjects design to varying degrees of synchronous or asynchronous presentations (either 366, 500, or 666ms). Infants who had been habituated to synchronous stimuli only demonstrated sensitivity to the largest difference between the two modalities the 666ms audio-lead. In contrast, infants who were habituated to the largest asynchronous presentation (666 ms) dishabituated to both the synchronous presentation and one in which the audio led the visual signal by 366ms. Together, these findings suggest that infants are sensitive to the temporal correlations of an audiovisual signal, however, they need to have the incongruencies highlighted and made salient before discriminating differences in timing.

Both the Pons et al (2012), and Lewkowicz (2010) studies establish that short-term experience may play a role in infant sensitivity to the temporal dynamics of multimodal signals. For both studies, highlighting the asynchronous nature of the stimuli served to alter discrimination patterns for both simple (e.g., ball bounces) and complex (e.g., the /ba/ speech syllable) displays. Both research groups suggest that sensitivity to these incongruent signals is

inconsistent and dependent on short-term perceptual experience. If short-term experience can reliably change infant sensitivity to cross-modal cues, are there characteristics of long-term experience that also might drive infant sensitivity to the spectral and temporal cues of the speech signal?

Shortly after birth, infants preferentially attend to face-like configurations (Goren, Sarty, & Wu, 1975), favor their mother's voice over a stranger's (DeCasper & Fifer, 1980), and demonstrate a left-dominant hemodynamic response for speech stimuli (Pena et al., 2003) similar to that of adults when perceiving language (for review, see Tervaniemi & Hugdahl, 2003). Investigations focusing on how infants process facial features have found that eyes are a highly salient social cue, in that younger infants preferentially attend to them above other facial features and cues until approximately the seventh month (Haith, Bergman, & Moore, 1977; Hunnius, & Geuze, 2004). Between 7- and 12-months of age, infants develop a variety of perceptual and productive abilities as they become experts in the constructs of their native language, canonically babbling between 6- and 7-months (Oller, Wieman, Doyle, & Ross,1975), producing language-specific phonemes and short words at around 10-months (Ferguson & Farwell, 1975), and undergoing perceptual narrowing for non-native speech contrasts between 9- and 12-months (Werker & Tees, 1984). In 2012, Lewkowicz and Hansen-Tift investigated whether these perceptual and production skills may be related to where infants attend when presented with an audiovisual speech paradigm. In an eye-tracking paradigm, infants from 4- to 10-months were presented with a video of a monolingual English speaker delivering short, infant appropriate salutations. Lewkowicz and Hansen-Tift predicted that if audiovisual speech perception was tied to developmental age and language abilities, than older infants (between 8- and 12-months) but not younger infants (between 4- and 8-months) would focus on the mouth of the speaker instead

of the eyes. They found that 4- and 6- month olds infants preferred to look at the eyes of the speaker, while attention to the mouth was correlated with age group for older infants, with selective attention to the mouth increasing as age increased. In a second experiment, they presented infants across the same age ranges with a Spanish audiovisual presentation and found that the results were mirrored and even more extreme as infant age increased. Younger infants still preferred to focus on the eyes of the speaker, while older infants spent even more time fixating on the mouth of the speaker. These results highlight how infants may be using the articulatory information in the visual signal to not only adequately perceive speech, but also to scaffold their own speech production.

The current study attempts to combine the approaches and findings previously reviewed and aid in filling the gaps about our understanding of audiovisual speech perception in the nascent language learner. Addressing the early work of audiovisual perception in young infants (e.g., Kuhl & Meltzoff, 1982; Kuhl & Meltzoff, 1984; Kuhl et al., 1991), the temporal binding window hypothesis (Hillock et al., 2011) and temporal saliency accounts (e.g., Lewkowicz, 2010; Pons et al., 2012), we will gauge whether infants are a) sensitive to audiovisual synchrony manipulations, b) depend upon temporal incongruencies being highlighted, and c) determine whether low-level temporal cues might provide additional information to the infant as they perceive audiovisual speech.

Aims and Hypotheses of the Current Study

The current study investigated infant sensitivity to asynchronous audiovisual presentations using more ecologically-valid stimuli than have typically been employed. In previous investigations, brief stimuli (e.g., tone-beeps or single phonemes) were used. But people actually speak to infants in multi-syllable utterances and sentences spanning more than a

second. We chose to use tri-syllable words to assess whether stimulus length might provide infants with more time to recognize the asynchronous signals and do so with audiovisual relationships common in their everyday life – those of words. We hypothesized that due to the longer presentation length and greater ecological validity of the stimuli, infants would look longer to asynchronous videos due to the novel nature of the audiovisual incongruencies.

The second aim of the study was to address how developmentally-based changes in infants' face and speech processing might help or hinder audiovisual speech perception. If in the last quarter of the first year, infants prefer to look at the mouth of a speaker (in-line with Lewkowicz & Hansen-Tift, 2012), then we predicted that they may be *more* sensitive to asynchronous manipulations due to their greater familiarity with the articulatory dynamics of acoustic speech and would look longer to asynchronous presentations. In contrast, younger infants may show the opposite effect as they are still becoming familiar with the knowledge that how the mouth moves affects the resulting acoustic signal.

Finally, the current study also sought to determine whether place of articulation influenced sensitivity to audiovisual speech. Articulatory dynamics that are more visibly accessible to the infant, such as labial consonants, may demonstrate a superiority effect to those that are produced further back in the mouth, such as alveolar and velar consonants. To investigate whether place of articulation and articulator visibility influenced infant sensitivity to the temporal congruency of the audiovisual speech signal, half of the infants were presented with a primarily labial tri-syllable word (*mufapi*) and the other half were exposed to a primarily alveolar tri-syllable word (*kalisue*). We predicted that infants would show greater sensitivity to the labial word over the alveolar word because the articulators are more visibly accessible during speech and would scaffold infants into recognizing the asynchronous audiovisual signal.

Unlike the typical IPLP, a single-screen display was used. This modification allowed us to insure that infants were capable of recognizing the differences between stimulus presentations (asynchronous vs. synchronous) because they did not need to discriminate between two screens but rather sequential presentations. Adult studies have found that audiovisual perception and integration may be gated by selective attention (e.g., Tippana, Andersen, & Sams, 2008) so it was imperative to provide the infants with presentations that did not divide their attention or cognitive resources. Looking preferences were determined by whether the infant was looking at the screen or away, with the latter indicating that they had grown bored of the display or failed to discriminate the differences. If during the second presentation (regardless of synchrony condition), infants were prone to looking away, it would suggest that they were failing to differentiate the differences between the two videos and perceived the presentations as being identical – in-line with a broad temporal binding window account. However, if infant attention was sustained during the second video, it would suggest that they had discriminated the differences between presentations and were sensitive to low-level temporal cues available in audiovisual speech perception.

Method

Participants

Twenty-six infants participated in the experiment. Five were excluded for either extreme fussiness ($N = 3$), having an ear infection at-test ($N = 1$), or having more than one ear infections since birth ($N = 1$). An additional infant was excluded due to a fire alarm occurring during the experimental session. The remaining 20 infants (11 boys, 9 girls) successfully completed the study. The age range was 169 – 295 days (5.5 – 9.7 months) and the average age was 219.5 days (7.2 months, $SD = 1.2$ months). Parental reports indicated that all infants were born full-term, had

a history of 1 or fewer ear infections, and American English input was greater than 80%. Participants were recruited by tracking birth announcements in area newspapers and sending request letters to their homes. Once the parents sent back their response card demonstrating their interest, they were contacted by telephone and an experiment session was scheduled. Recruitment was limited to the New England area due to travel restrictions, resulting in a primarily Caucasian and college-educated sample.

Materials

The synchronous speech videos were composed of brief clips of a woman articulating either *kalisue* (the primarily alveolar word) or *mufapi* (the primarily labial word) in engaging, but not infant-directed, speech. The stimuli were identical to those used in Baart, Vroomen, Shaw, and Bortfeld (2014; see for initial stimulus creation details). Each clip was ~1.2 seconds long. For the synchronous video conditions, 18 identical clips (either kalisue or mufapi, depending on word group) were presented sequentially with 500ms inter-stimulus intervals (ISIs).

For the asynchronous speech videos, the clips were altered using Adobe Premiere Pro CS6 so the auditory signal (spoken kalisue or mufapi) preceded visual onset by 300ms. During the auditory-only presentation, a blank screen accompanied the sound. For the asynchronous video conditions, 18 identical clips were also presented sequentially with 500ms inter-stimulus intervals, see Figure 1 for a schematic of trial presentation.

For both synchronous and asynchronous videos, infants were first presented with a silent articulating face to alert them that visual information would be provided on the screen. Infants in the kalisue word group were presented with silently articulated mufapi and infants in the mufapi word group were presented with silently articulated kalisue.

Procedure

Stimuli were presented on a standard computer monitor in a quiet experimental room. Infants sat on their parents' laps and external distractions (e.g., toys, other people) were removed from the room. Parents were instructed to allow their child to watch the video naturally, without trying to engage them with the presentation or engage them interpersonally. During the experiment, parents were equipped with headphones playing instrumental music to mask the auditory stimuli and reduce the likelihood of potentially influencing their infant. After the experiment, parents were asked what they had heard and only one reported being able to make out the sound and suggested it was "macaroni."

Infants were randomly assigned to experimental group conditions with word group as a between-subjects factor. Half of the participants were assigned to see kalisue for both synchronous and asynchronous presentations and the remaining infants saw mufapi for both synchronous and asynchronous presentations. Video order was counter-balanced across participants and word groups, with half of the participants seeing the synchronous video first and then the asynchronous video second and the order was reversed for the remaining participants. In between videos, an experimenter stepped into the room, muted the headphones, and asked the parents if they or their child needed a break. Due to the short duration of each video (a little over one-minute), all parents opted not to take a break.

During set-up and stimulus presentation, infant looking patterns were video-recorded for off-line analyses.

Coding Analyses

Videos were coded frame-by-frame (29.97/sec) using Adobe Premiere Pro CS6 software. For synchronous videos, the first frame where the child was provided with a visual stimulus was considered the start of the trial and the trial ended after the auditory word completed. For

asynchronous videos, the first frame where the child was provided with an auditory stimulus was considered the start of the trial and the trial ended 300ms after auditory word completion (to account for continued visual information due to audio-lead). A criterion of one-second was employed for determining whether a child was looking at the screen (*on*) or looking away (*off*). Each participant video was coded off-line by two independent coders blind to the hypotheses. In cases of disagreement, a third coder was brought in. Inter-rater reliability was greater than 90%. After checking reliability, total number of frames on and total number of frames off were converted to time units (1 frame = 33.336 ms) and proportion of looking time on was calculated by taking the total amount of looking time on and dividing it by the sum of total amount of looking time on and total amount of looking time off. All infants included in final analyses had greater than 50% proportion of total looking time on for both synchronous and asynchronous videos.

Results

Infants were median-split by age (younger vs. older) to ensure equal sample sizes in both age groups. For the younger age group, the age range was 169 – 211 days (5.55 – 6.93 months; M_{age} = 6.23 months). For the older age group, the age range was 218 – 295 days (7.2 – 9.7 months; M_{age} = 8.2 months).

Overall, infants were quite engaged, with the proportion of looking time on (PLT), regardless of age or word, being 0.93 (*SD* = 0.11). No differences were found between video orders, with infants looking at the second proportion of looking time for the second video being 0.92 (*SD* = 0.04). As a result, video order was collapsed for the remaining analyses.

A 2(age: younger, older) x 2(word: kalisue, mufapi) repeated measures analysis of variance (ANOVA) was conducted to determine whether PLT to synchronous stimuli

significantly differed from PLT to asynchronous stimuli as a result of word condition or age. There was a main effect of word, with both age groups of infants preferring to look at mufapi presentations regardless of whether it was synchronous or asynchronous, $F(1, 16) = 5.010$, $p < .05$, see Figure 2. There was a marginally significant main effect of look, with asynchronous presentations being preferentially attended to over synchronous presentations, $F(1, 16) = 4.005$, $p = .063$, see Figure 3. In addition, there was a marginally significant main effect of age, with younger infants looking longer to both presentations than older infants, $F(1, 16) = 3.522$, $p = .079$, see Figure 4. Planned t-tests demonstrated a marginally significant effect of age and presentation, with older infants, but not younger infants, looking longer to asynchronous presentations than synchronous presentations, $t(9) = -2.220$, $p = .054$, see Figure 4.

Discussion

The current study sought to investigate whether infants are sensitive to audiovisual temporal asynchrony in speech and if sensitivity was related to chronological age and, thus, experience with the natural correlations of the auditory and visual speech signals. In addition, we were interested in seeing whether articulatory phonetics, or how accessible the visual speech cues are, would help or hinder sensitivity to temporal incongruencies. We hypothesized that a) due to the longer presentation length and greater ecological validity of the stimuli, infants would look longer to asynchronous videos due to the novel nature of the audiovisual incongruencies, b) older infants would be *more* sensitive to asynchronous manipulations due to their greater familiarity with the articulatory dynamics of acoustic speech and would look longer to asynchronous presentations, and c) that infants would show greater sensitivity to the labial word over the alveolar word because the articulators are more visibly accessible during speech and would scaffold infants into recognizing the asynchronous audiovisual signal. Overall, our third

hypothesis was supported, with infants preferring to look at labial words over alveolar words, suggesting that the increased accessibility to the visual articulators helped in recognizing differences between synchrony conditions. The other two hypotheses were only moderately supported. We discuss our findings and ramifications of the work below.

The Role of Articulatory Visibility in Sensitivity to Asynchronous Audiovisual Speech

Indeed, we found that place of articulation and accessibility to the visual cues of the articulators was influential in how engaged infants were with the presentations. Older infants, in particular, preferred to look at asynchronous labial presentations, suggesting that they were sensitive to the temporal incongruencies and interested in how the asynchrony may be occurring. Younger infants, however, performed near-ceiling for both synchronous and asynchronous presentations. One might purport that this shows they also were sensitive to the temporal inconsistencies and were re-engaged with the second video, however, it is more likely that they failed to notice the differences between presentations and were still processing the complex stimuli, in-line with developmental theories of stimulus complexity, recognition memory, and visual gaze (Aslin & Fiser, 2005; Rose, Feldman, & Jankowski, 2004).

Developmental Age as a (Potential) Moderator of Sensitivity to Spectrotemporal Relationships in Speech

Older infants moderately preferred to look at asynchronous stimuli, regardless of video order, a surprising finding considering that the adult literature (e.g., Hillock et al., 2011) would suggest that they would erroneously bind the multimodal signals into a single percept due to the temporal binding window hypothesis. In addition, this finding contrasts to that of Lewkowicz (2010) and Pons et al. (2012), who would assert that asynchronous presentations would need to be presented first for infants to be sensitive to the temporal manipulations. One strength of the

current investigation is that it mirrors audiovisual relationships inherent in the infant's environment with tri-syllable words that allow more time to process the asynchrony but also are more reflective of the infant's language experience. In comparison, younger infants did not demonstrate a preference for visual presentation, looking nearly at ceiling for both videos. These findings are somewhat in-line with those of Lewkowicz and Hansen-Tift (2012) but extend them further – face processing mechanisms involved in speech perception developmentally shift as a reflection of infant language production, but as our study suggests, also influences discriminatory abilities for low-level cues the infant may utilize when learning the relationship between the visual articulators and acoustic signal.

Ecological Validity and Predictive Value Impact Sensitivity to Audiovisual Asynchrony

Finally, there is moderate evidence that infants preferred to look at asynchronous presentations overall (despite older infants primarily contributing to this effect). Although speculations need to be restrained, we suggest that this sensitivity to temporal incongruencies in the speech signal are driven by the ecological validity of our stimuli and the predictive information available to infants. Previous audiovisual studies have found that the degree to which the visual stimulus is predictive of the auditory stimulus can influence speed and efficiency of audiovisual perception (Kopp & Dietrich, 2013; Stekelenburg & Vroomen, 2007). In the current study, the primarily labial and alveolar words differ in the variety of phonemes or sounds they can precede. Bilabial closures are relatively limited to /b/, /p/, /m/, thus infants are likely to be sensitive to the three-to-one mapping, particularly because the mufapi token consists of 2/3 of the possible phonemes. Alveolar consonants in the English language are much more varied, with /n/, /t/, /d/, /s/, /r/, /l/, and /z/ associated with this place of articulation. In this case,

there is a lower level of predictability for the acoustic signal following, due to both visible information and a many-to-one mapping problem.

Limitations and future directions

Our results highlight how visible accessibility of the articulators and developmental experience can aid in infants recognizing asynchronous relationships and contradict previous investigations suggesting that their temporal binding windows are relatively broad and that the incongruencies in the signals must be emphasized to boost infant discrimination. Although we are quite confident that these results are due to the strength of our experimental manipulations and the ecological validity of our stimuli, a few caveats must be acknowledged and developed in future lines of inquiry.

One precaution to recognize is that the sample size was relatively small, with only ten infants in each age group. Although this is not necessarily uncommon in the developmental literature, the marginally significant effects and innovative findings should be taken with a grain of salt until the study has been extended or replicated with a larger number of participants. In addition, we can only address one of the key concerns of our literature review – that suggested by Kuhl and colleagues (1982; 1984; 1991) that audiovisual speech perception is not reliant on temporal correlations but rather phonetic cues. While we cannot directly respond to this argument based on the current findings, we have demonstrated that infants are sensitive to these multimodal temporal consistencies and that they are reliable cues when perceiving speech. To justify our conclusions and address the assertions of Kuhl and colleagues, we have begun work utilizing sine wave speech similar to Baart et al. (2014) to examine infant sensitivity to audiovisual speech synchrony when spectral and phonetic cues are degraded, forcing infants to rely considerably on temporal information to identify audiovisual incongruencies.

In our current discussion of our findings, we concede that the younger infant looking patterns appear to be less influenced by experimental manipulations and more by developmental shifts in visual gaze to complex stimuli. Across synchronous and asynchronous conditions, younger infants looked approximately 95% of the time, a finding that could reflect sensitivity to the video conditions being different but is more likely to be a ceiling effect. We cannot assert based on our current looking data which conclusion is most appropriate, but additional methodologies including electrophysiological and neurophysiological indices of processing may aid in delineating these possibilities.

For instance, neuroimaging studies have suggested that the posterior superior temporal sulcus (pSTS) is a neural hub of multisensory integration, particularly for speech (Kreifelts, Ethofer, Shiozawa, Grodd, & Wildgruber, 2009; Nath, Fava, & Beauchamp, 2011; Stevenson, VanDerKlok, Pisoni, & James, 2011). Children and adult McGurk perceivers have been found to have greater activation associated with the pSTS when experiencing a McGurk illusory percept, while non-perceivers have had little-to-no hemodynamic signal change in the same region during McGurk-inducing presentations (Nath et al., 2011; Nath & Beauchamp, 2012). McGurk perceivers have also been shown to have their behavioral responses and subsequent pSTS activation responses attenuated when experiencing transcranial magnetic stimulation (TMS) applied to the area (Beauchamp, 2010). Finally, pSTS activation has been shown to be modulated by the temporal congruencies between auditory and visual signals (Noesselt et al., 2007) and whether someone is in speech mode while perceiving sine wave speech (Mottonen et al., 2006). To better understand infant audiovisual perception and differentiate between looking patterns and actually discriminatory abilities, it is imperative that future work include additional methodologies that investigate the role and development of the pSTS in infancy.

Conclusion

The current study demonstrates that infant audiovisual speech perception is more robust and reliant on amodal cues than previously anticipated. Unlike Kuhl and colleagues (1982, 1984, and 1991) and Patterson and Werker (2003), we found that infants *are* sensitive to the temporal relationships between the auditory and visual speech signals when perceiving speech and that these low-level cues may be informative to the young perceiver. Also, in contrast to both adult (e.g., Hillock et al., 2011) and infant (e.g., Lewkowicz, 2010; Pons et al., 2012) work investigating infant sensitivity to asynchrony, we found that older infants in particular fail to erroneously bind asynchronous audiovisual information regardless of presentation order, suggesting that their temporal binding windows are relatively narrow for speech and saliency of the incongruencies between modalities is inconsequential for perception. We attribute this finding to the greater ecological validity of our stimuli and the comparison of highly predictive visual cues to cues with low predictability (e.g., labial vs. alveolar).

Overall, our work suggests that infants *are* sensitive to the amodal temporal correlations spanning both the visual and acoustic speech signals and that this development is driven by familiarity and developmental shifts in face processing, the role of multimodal predictability of low-level cues, and the interaction between experience and sensitivity to audiovisual speech relationships.

References

Aslin, R. N., & Fiser, J. (2005). Methodological challenges for understanding cognitive development in infants. *TRENDS in Cognitive Sciences,* 9(3), 92-98.

Baart, M., Vroomen, J., Shaw, K., & Bortfeld, H. (2014). Degrading phonetic information affects matching of audiovisual speech in adults, but not infants. *Cognition, 130*(1), 31-43.

Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). FMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk Effect. *The Journal of Neuroscience, 30*(7), 2414-2417.

Bergelson, E. & Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common words. *Proceedings of the National Academy of Sciences of the United States of America, 109*(9), 3253-3258.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology,5*(7), e1000436.

DeCasper, A. J., & Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science, 208,* 1174-1176.

Desjardins, R. N., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology, 45,* 187-203.

Eskelund, K., Tuomainen, J., & Andersen, T. S. (2011). Multistage audiovisual integration of speech: dissociating identification and detection. *Experimental Brain Research, 208*(3), 447-457.

Falck-Ytter, T., Bakker, C., & von Hofsten, C. (2011). Human infants orient to biological motion rather than audiovisual synchrony. *Neuropsychologia, 49*(7), 2131-2135.

Ferguson, C. A., & Farwell, C. B. (1975). Words and sounds in early language acquisition. *Language, 51*(2), 419-439.

Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., & Gordon, L. (1987). The eyes have it: Lexical and semantic comprehension in a new paradigm. *Journal of Child Language, 14*(1), 23-45.

Goren, C. C., Sarty, M., & Wu, P. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics, 56*(4), 544-549.

Haith, M. M., Bergman, T., & Moore, M. J. (1977). Eye contact and face scanning in early infancy. *Science, 198*(4319), 853-855.

Hillock, A. R., Powers, & Wallace, M. T. (2011). Binding of sights and sounds: Age-related changes in multisensonsory temporal processing. *Neuopsychologia, 49*, 461-467.

Hunnius, S., & Geuze, R. H. (2004). Developmental changes in visual scanning of dynamic faces and abstract stimuli in infants: A longitudinal study. *Infancy, 6*(2), 231-255.

Innes-Brown, H., Barutchu, A., Shivdasani, M. N., Crewther, D. P., Grayden, D. B., & Paolini, A. (2011). Susceptibility to the flash-beep illusion is increased in children compared to adults. *Developmental Science, 14*(5), 1089-1099.

Kopp, F., & Dietrich, C. (2013). Neural dynamics of audiovisual synchrony and asynchrony perception in 6-month-old infants. *Frontiers in Psychology, 4*, 1-13.

Kreifelts, B., Ethofer, T., Shiozawa, T., Grodd, W., & Wildgruber, D. (2009). Cerebral representation of non-verbal emotional perception: fMRI reveals audiovisual integration area between voice- and face-sensitive regions in the superior temporal sulcus. *Neuropsychologia, 47*, 3059-3066.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science, 218*(4577), 1138-1141.

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development, 7*(3), 361-381.

Kuhl, P. K., Williams, K. A., Meltzoff, A. N. (1991). Cross-modal speech perception in adults and infants using nonspeech auditory stimuli. *Journal of Experimental Psychology: Human Perception and Performance, 17*(3), 829-840.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255*(5044), 606-608.

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. Nature Reviews *Neuroscience, 5*, 831-843.

Lewkowicz, D.J. (2010). Infant perception of audio-visual speech synchrony. *Developmental Psychology, 46*, 66-77.

Lewkowicz, D., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning language. *Proceedings of National Academy of Sciences, 109*(5), 1431-1436.

Maye, J., Werker, J. F., & Gerken, L. A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82,* B101-B111.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

Mottonen, R., Calvert, G. A., Jaaskelainen, I. P., Matthews, P. M., Thesen, T., Tuomainen, J., & Sams, M. (2006). Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. *NeuroImage, 30*, 563-569.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language,17*(2), 357-374.

Nath, A. R., & Beauchamp, M. S. (2011). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *NeuroImage, 59*(1), 781-787.

Nath, A. R, Fava, E. E., & Beauchamp, M. S. (2011). Neural correlates of interindividual differences in children's audiovisual speech perception. *The Journal of Neuroscience, 31*, 13963-13971.

Noesselt, T., Rieger, J. W., Schoenfeld, M. A., Kanowski, M., Hinrichs, H., Heinze, H. –J., & Driver, J. (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *The Journal of Neuroscience, 27*(42), 11431-11441.

Oller, D. K., Wieman, L. A., Doyle, W. J., & Ross, C. (1975). Infant babbling and speech. *Journal of Child Language, 3*, 1-11.

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science, 6*(2), 191-196.

Pena, M., Maki, A., Kovacic, D., Dehaene-Lambertz, G., Koizumi, H., Bouquet, F., & Mehler, J. (2003). Sounds and silence: an optical topography study of language recognition at birth. *Proceedings of the National Academy of Sciences of the United States of America, 100*(20), 11702-11705.

Pena, M., Mehler, J., & Nespor, M. (2011). The role of audiovisual processing in early conceptual development. *Psychological Science, 22*(11), 1419-1421.

Pons, F., Teixido, M., Garcia-Morera, J., & Navarra, J. (2012). Short-term experience increases infants' sensitivity to audiovisual asynchrony. *Infant Behavior & Development, 35*, 815-818.

Pruden, S. M., Hirsh-Pasek, K., Golinkoff, R., & Hennon, E. A. (2006). The Birth of Words: Ten-Month-Olds Learn Words Through Perceptual Salience. *Child Development, 77*(2), 266-280.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212*, 947-949.

Rose, S. A., Feldman, J. F., & Jankowski, J. J. (2009). Information processing in toddlers: Continuity from infancy and persistence of preterm deficits. *Intelligence, 37*(3), 311-320.

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience, 19*(12), 1964-1973.

Stevenson, R. A., VanDerKlok, R. M., Pisoni, D. B., & James, T. W. (2011). Discrete neural substrates underlie complementary audiovisual speech integration processes. *NeuroImage, 55,* 1339-1345.

Tervaniemi, M., & Hugdahl, K. (2003). Lateralization of auditory-cortex functions. *Brain Research Review, 43*, 231-246/

Tippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology, 16*(3), 457-472.

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition, 96*(1), B13-22.

Vroomen, J., & Baart, M. (2009). Phonetic recalibration only occurs in speech mode. *Cognition, 110*(2), 254-259.

Vroomen, J., & Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition, 118*(1), 75-83.

Watanabe, H., Homae, F., Nakano, T., Tsuzuki, D., Enkhtur, L., Nemoto, K., Dan, I., & Taga, G. (2013). Effect of auditory input on activation in infant diverse cortical regions during audiovisual processing. *Human Brain Mapping, 34*, 543-565.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*, 49-63.

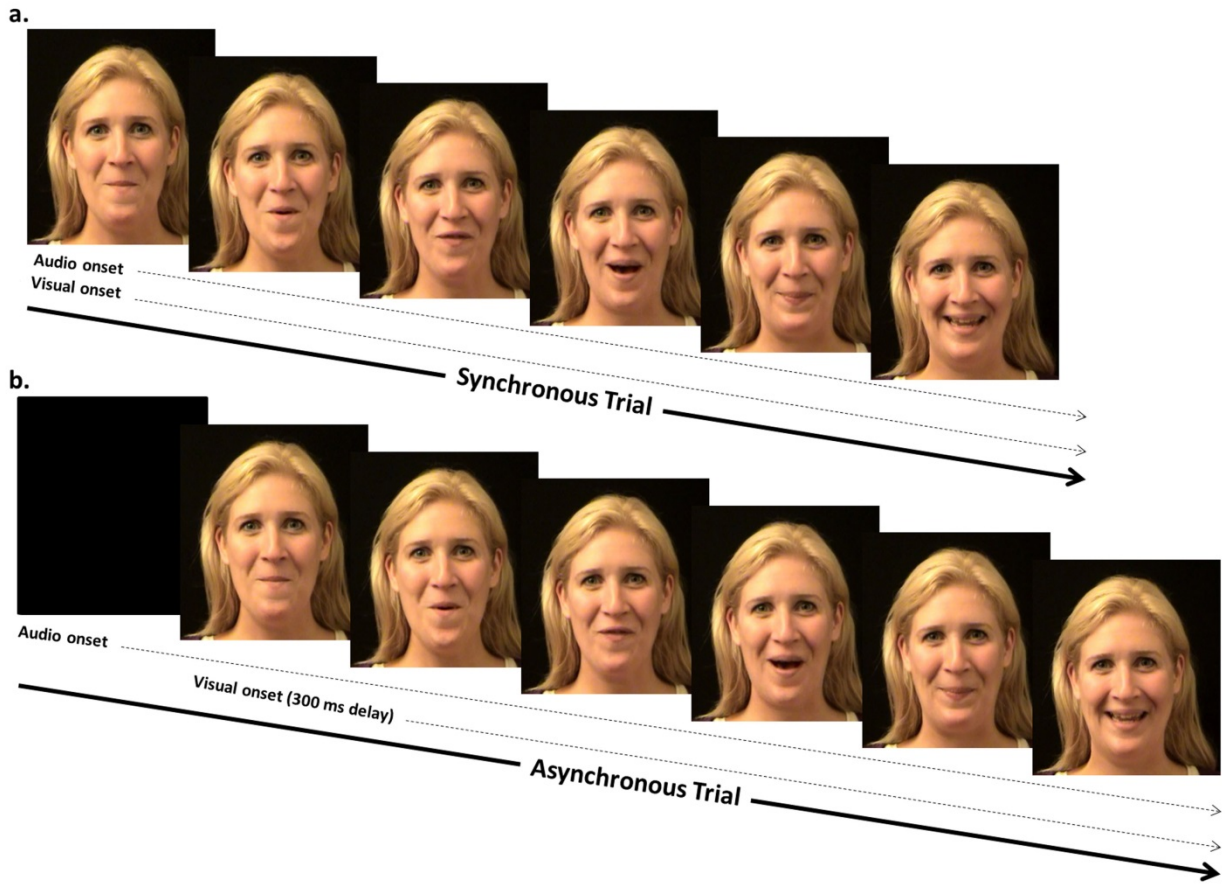

Figure 1. Schematic of single trial for either the a) synchronous video or b) asynchronous video.

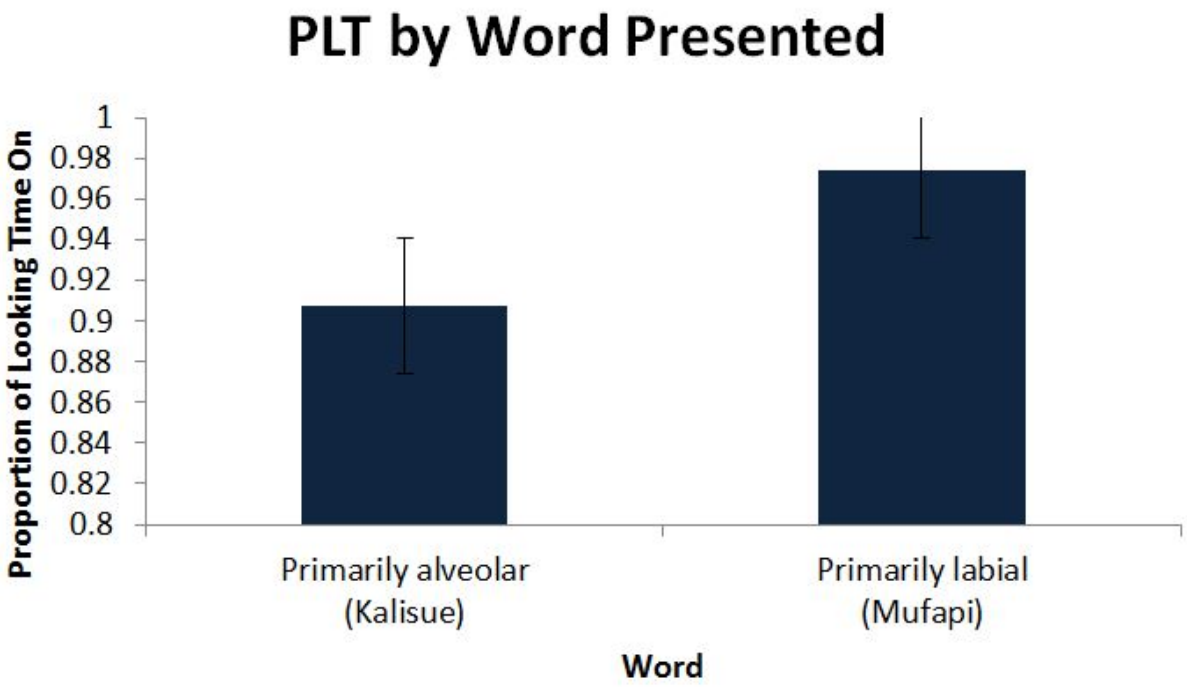

Figure 2. Average proportion of looking time to primarily alveolar and primarily labial words.

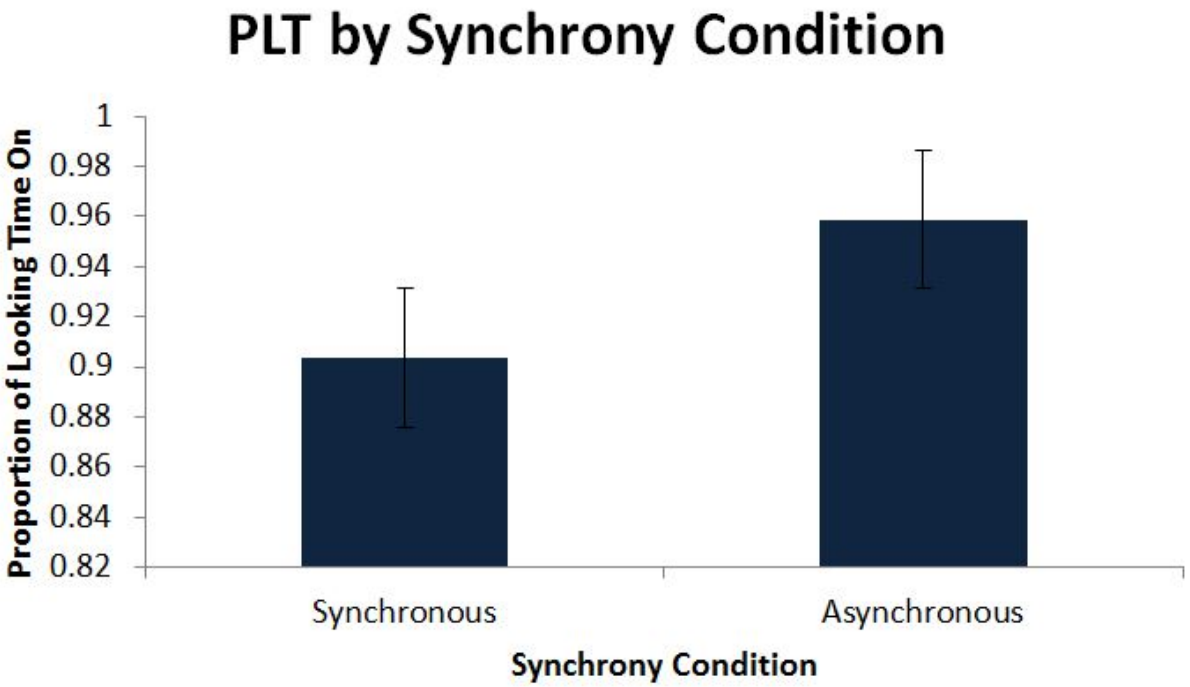

Figure 3. Average proportion of looking time to synchronous and asynchronous video presentations, regardless of video order.

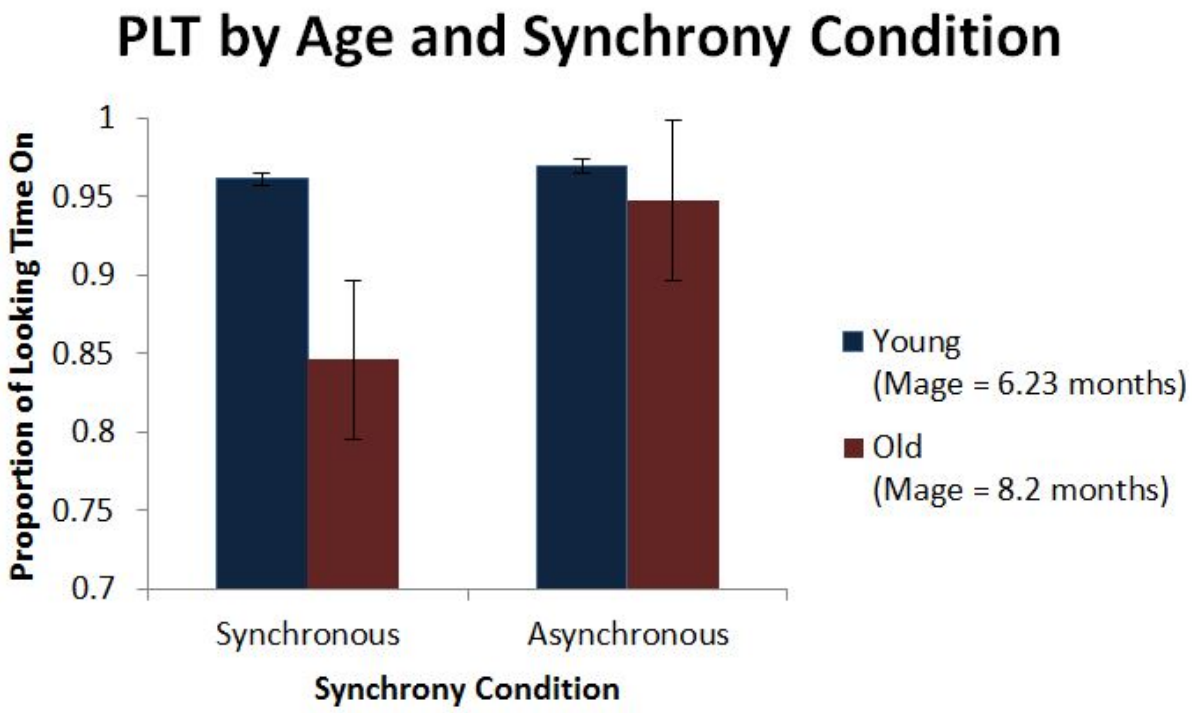

Figure 4. Average proportion of looking time to synchronous and asynchronous video presentations separated by infant age group.